# 1 前回に引き続き

## 1.1 $L_1$ Regularized Logistic Regression

次の教師付き学習問題を考え：Training set は $\{(x_i, y_i),\ i = 1, \ldots, M\}$, $x_i \in \mathbb{R}^N$, $y_i \in \{-1, 1\}$, ロジスティック回帰モデル（Logistic Regression Model）の確率分布は：

$$p(y|x;\theta) = \sigma(y\theta^T x) = \frac{1}{1+\exp(-y\theta^T x)} \quad , y \in \{-1, 1\} \tag{1.1}$$

ただし，$\theta \in \mathbb{R}^N$ はロジスティック回帰モデルのパラメーターである．
ラプラス分布の事前確率は

$$p(\theta) = \prod_j^N \bar{p}(\theta_j) = \prod_j^N \frac{1}{2\beta} \exp\Big(-\frac{|\theta_j|}{\beta}\Big) = \Big(\frac{1}{2\beta}\Big)^N \exp\Big(-\frac{\|\theta\|_1}{\beta}\Big),\ \beta > 0 \tag{1.2}$$

パラメーター $\theta$ に対しての最大事後確率（maximum a posteriori）推定は

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^M -\log p(y_i|x_i;\theta) + \frac{\|\theta\|_1}{\beta} \tag{1.3}$$

であり，次の式と同値

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^M -\log p(y_i|x_i;\theta),\ \text{subject to } \|\theta\|_1 \le C \tag{1.4}$$

# 2 正則化付き最小化問題

$$\underset{w\in\mathbb{R}^n}{\text{minimize}} \quad \phi_\gamma(w) := f(w) + \gamma r(w) \tag{2.1}$$

近年，式 2.1 の正則化付き最小化問題はかなり研究されていて，勾配法または劣勾配法のような汎用アルゴリズムは幾つがある．一般的な関数に対して収束率がせいぜい $\frac{1}{k^2}$ すぎないが，関数の特徴を利用すればもっと早く収束できる．

# 3 First-Order Methods

## 3.1 準備

**定義 3.1.** 関数 $f$ の定義域 $\mathbf{dom}\, f$ が凸であり，かつ，次の条件

$$\forall \alpha \in [0,1], f(\alpha x + (1-\alpha)y) + \mu\frac{\alpha(1-\alpha)}{2}\|x-y\|^2 \le \alpha f(x) + (1-\alpha)f(y) \tag{3.1}$$

を満たすとき，$f$ を **$\mu$-強凸関数 ($\mu$-strongly convex function)** という．

- *$\mu$-strongly convex function* の一次条件：$f$ が一階微分可能，かつ

$$\forall x, y \in \mathbf{dom}\, f,\quad f(y) \geq f(x) + \langle \nabla f(x), x - y\rangle + \frac{\mu}{2}\|x - y\|^2 \tag{3.2}$$

- *$\mu$-strongly convex function* の二次条件：$f$ が二階微分可能，かつ

$$\forall x, y \in \mathbf{dom}\, f,\quad \langle \nabla^2 f(x)y, y\rangle \geq \mu\|y\|^2 \tag{3.3}$$

特に，$\|\cdot\| = \|\cdot\|_2$ の場合，$\nabla^2 f(x) \succeq \mu\mathbb{I}$

**定義 3.2.** 関数 $f$ が

$$\exists L \geq 0, \|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \tag{3.4}$$

を満たすとき，$f$ は**リプシッツ連続 (Lipschitz continuity)** である．$f$ がリプシッツ条件を満たすための最小の $L$ の値を $f$ の**リプシッツ定数 (Lipschitz constant)** という．

リプシッツ連続は関数における通常の連続よりも強い平滑 *(smoothness)* 条件である．直観的には，リプシッツ連続は変化の速度を制限するものである．

**命題 3.3.** 微分可能かつ凸のリプシッツ連続関数 $f$ を $f \in C_L^{1,1}$ と記述する．$f$ は二階微分可能ならば，

$$f \in C_L^{1,1} \Leftrightarrow \langle \nabla^2 f(x)y, y\rangle \preceq L\|y\|^2 \tag{3.5}$$

$C_L^{k,p}$ とは $k$ 階連続微分可能（$k$ *times continuously differentiable*）かつ $p$ 階微分が $L$ リシッツ連続の関数の集合.

**命題 3.4.** 連続微分可能の *$\mu$-strongly convex* 関数集合を $S_\mu^1$ という，また，$f$ が *$\mu$-strongly convex function* かつ $f \in C_L^{k,p}$ の関数 $f$ の集合は $S_{\mu,L}^{k,p}$ という．

**命題 3.5.** 一般的には，平滑かつ凸 *(smooth and convex)* 関数 $f$ に対して，

$$\exists 0 \leq \mu \leq L,\ \forall x,\ \mu\mathbb{I} \preceq \nabla^2 f(x) \preceq L\mathbb{I} \tag{3.6}$$

特に，凸関数 $f = \frac{1}{2}x^T Ax$ に対して，$\mu\mathbb{I} \preceq A \preceq L\mathbb{I}$

**命題 3.6.**

$$f \in C_L^{1,1} \Rightarrow f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 \tag{3.7}$$

証明.

$$\begin{aligned}
&f(y) - f(x) - \langle \nabla f(x), y - x\rangle \\
=&\int_0^1 \langle \nabla f(x + \tau(y - x)), y - x\rangle d\tau - \int_0^1 \langle \nabla f(x), y - x\rangle d\tau \\
=&\int_0^1 \langle \nabla f\big(x + \tau(y - x)\big) - \nabla f(x), y - x\rangle d\tau \\
\leq&\int_0^1 \|\nabla f\big(x + \tau(y - x)\big) - \nabla f(x)\|_* \cdot \|y - x\| d\tau \\
\leq&\int_0^1 L\|\tau(y - x)\| \cdot \|y - x\| d\tau \\
=&\frac{L}{2}\|y - x\|^2
\end{aligned} \tag{3.8}$$

□

**命題 3.7.** すべての $f \in S_{\mu,L}^{1,1}$ に対して,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2 \tag{3.9}$$

証明. 本 [2] の 66 ページを参照 □

**定義 3.8.** 数列 $\{\delta_i\}$ が 0 に収束することを仮定すると，**収束率（rate of convergence）**は

$$\lim_{i \to \infty} \frac{\delta_{i+1}}{\delta_i} = \sigma \tag{3.10}$$

に定義する.

- $\sigma = 0$ : 超線形収束（*superlinear rate*）
- $\sigma = 1$ : 劣線形収束（*sublinear rate*）
- $\sigma \in (0, 1)$ : 線形収束（*linear rate*）

## 3.2 勾配法（Gradient method）

それでは，勾配法の収束率を求めよう．次の問題を考え : 関数 $f \in C_L^{1,1}$，$x_{k+1} = x_k - h_k \nabla f(x_k)$, その最小値 $\min f(x)$ を求める.
命題 3.6 より,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\| \\ &= f(x_k) - h_k \|\nabla f(x_k)\|^2 + \frac{h_k^2}{2} L \|\nabla f(x_k)\|^2 \\ &= f(x_k) - h_k (1 - \frac{h_k}{2} L) \|\nabla f(x_k)\|^2 \end{aligned} \tag{3.11}$$

定ステップ法（constant step strategy）を用いて，$h_k = h$．最適なステップサイズを得るため,

$$\Delta(h) = -h(1 - \frac{hL}{2}) \tag{3.12}$$

の最小値 $\min \Delta(h)$ を求める．$\Delta'(h) = hL - 1 = 0$ ので，最適解 $h^* = \frac{1}{L}$ である．また，最適解 $x^*$ を取るときの最適値 $f^* = f(x^*) = \inf_x f(x)$ とおく.

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \frac{h}{2} \|\nabla f(x)\|_2^2 \\ &\leq f^* + \nabla f(x)^T (x_k - x^*) - \frac{h}{2} \|\nabla f(x)\|_2^2 \\ &= f^* + \frac{1}{2h} \Big( \|x_k - x^*\|_2^2 - \|x_k - x^* - h \nabla f(x_k)\|_2^2 \Big) \\ &= f^* + \frac{1}{2h} \Big( \|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 \Big) \\ &= f^* + \frac{L}{2} \Big( \|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 \Big) \end{aligned} \tag{3.13}$$

$f(x_{k+1})$ と $f^*$ の差を $\delta_{k+1} := f(x_{k+1}) - f^*$ とおくと，

$$\begin{aligned} 0 &\leq \frac{L}{2}\|x_{k+1} - x^*\|_2^2 \leq -\delta_{k+1} + \frac{L}{2}\|x_k - x^*\|_2^2 \\ &\leq \ldots \leq -\sum_{i=1}^{k+1}\delta_i + \frac{L}{2}\|x_0 - x^*\|_2^2 \\ &\leq -(k+1)\delta_{k+1} + \frac{L}{2}\|x_0 - x^*\|_2^2 \end{aligned} \tag{3.14}$$

変形すると，

$$\delta_{k+1} \leq \frac{L\|x_0 - x^*\|_2^2}{2(k+1)} \tag{3.15}$$

従って，$f \in C_L^{1,1}$ の時，$f$ は $\frac{1}{k}$ の劣線形収束率で収束する．が，もう少しスピードアップしたい．

### 3.3 Strongly Convex Function

$f \in S_{\mu,L}^{1,1}$ の時，もっと早いスピードで収束できる．式 3.9 より，

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &= \|x_k - h\nabla f(x_k) - x^*\|_2^2 \\ &= \|x_k - x^*\|_2^2 - 2h\nabla f(x_k)^T(x_k - x^*) + h^2\|\nabla f(x_k)\|_2^2 \\ &= \|x_k - x^*\|_2^2 + h^2\|\nabla f(x_k)\|_2^2 - 2h\big(\nabla f(x_k) - \nabla f(x^*)\big)^T(x_k - x^*) \\ &\leq \|x_k - x^*\|_2^2 + h^2\|\nabla f(x_k)\|_2^2 - 2h\big(\frac{\mu L}{\mu + L}\|x_k - x^*\|_2^2 + \frac{1}{\mu + L}\|\nabla f(x_k)\|_2^2\big) \\ &= (1 - h\frac{2\mu L}{\mu + L})\|x_k - x^*\|_2^2 + h(h - \frac{2}{\mu + L})\|\nabla f(x_k)\|_2^2 \\ &\leq (1 - h\frac{2\mu L}{\mu + L})\|x_k - x^*\|_2^2 \end{aligned} \tag{3.16}$$

$h(h - \frac{2}{\mu+L})$ を最小にするのステップサイズ $h = h^* = \frac{2}{\mu+L}$ とおくと，式 3.14 のように変形すれば，

$$\delta_{k+1} = f(x_k) - f^* \leq \frac{L}{2}\Big(\frac{L - \mu}{L + \mu}\Big)^{2k}\|x_0 - x^*\|_2^2 \tag{3.17}$$

従って，$f \in S_{\mu,L}^{1,1}$ の時，$f$ は線形収束率で収束できる．

## 4 Shrinking/Thresholding for Regularized Optimization

一般的には，式 2.1 の正則化関数 $r(w)$ はスムーズではなくてシンプルの関数（nonsmooth but simple）である．**Iterative shrinking/thresholding algorithm（IST）**または **Forward-backward splitting** とは，$f$ を対角二次ヘッセ行列と入れ替えて元問題をより簡単に解けるアルゴリズムである：

$$\begin{aligned} &\min_w \frac{1}{2\alpha}\|w - (x - \alpha\nabla f(w))\|_2^2 + \gamma r(w) \\ \Leftrightarrow \quad &\min_w \nabla f(w)^T(w - x) + \frac{1}{2\alpha}\|w - x\|_2^2 + \gamma r(w) \end{aligned} \tag{4.1}$$

**縮小オペレーター（shrinking operator/proximity operator/thresholding function）**を

$$S_\gamma(y,\alpha) := \arg\min_w \frac{1}{2\alpha}\|w-y\|_2^2 + \gamma r(w) \tag{4.2}$$

に定義すると．次のような反復アルゴリズムを考えることができる．

$$\begin{aligned} w_{k+1} &:= S_\gamma(w_k - \alpha_k \nabla f(w_k), \alpha_k) \\ &:= \arg\min_w (w-w_k)^T \nabla f(w_k) + \frac{1}{2\alpha}\|w-w_k\|_2^2 + \gamma r(w) \end{aligned} \tag{4.3}$$

一部の正則化関数に対して，子問題は簡単関数である．例えば，$r(w) = \|w\|_1$ の場合，

$$S_\gamma(y,\alpha) = \text{sign}\,(y)\max(|y|-\alpha\gamma, 0) \tag{4.4}$$

式 4.4 は Soft-thresholding function と呼ばれる．IST の利点は 1 反復あたりの計算量（勾配の計算コストと $S_\gamma(y,\alpha)$ の計算コスト）が小さいことである．一方，IST はステップサイズ $\alpha$ の選択方法が難しいことなどの問題点がある．

# 5 Second-order methods

申し訳ないが，Second-order methods わからない．略

# 6 Regularized dual averaging

Regularized dual averaging（RDA）は反復ごとに劣勾配を解くことだけではなく，すべての正則化項（Regularized items）を含む簡単な最適化問題を解くアルゴリズムである．

Regularized stochastic learning problem とは

$$\underset{w}{\text{minimize}}\, \phi_\gamma(w) := E_\xi f(w,\xi) + \gamma r(w) \tag{6.1}$$

$\xi$ は入力・出力ペアの分布に従う変数である．反復 $k$ ではランダムで $\xi$ 分布から $\xi_k$ を選び．$g_k \in \partial f(w_k, \xi_k)$ とおくと，平均劣勾配（averaged subgradient）を

$$\bar{g}_k = \frac{1}{k}\sum_{i=1}^k g_i \tag{6.2}$$

に定義できる．そして，次の子問題を解く：

$$w_{k+1} := \arg\min_w \bar{g}_k^T w + \gamma r(w) + \frac{\alpha}{\sqrt{k}}\|w-w_0\|^2 \tag{6.3}$$

その平均収束率 $\bar{w}_k$ は

$$E_{\phi_\gamma}(\bar{w}_k) - \phi_\gamma^* \le \frac{C}{\sqrt{k}} \tag{6.4}$$

# 参考文献

[1] Efficient L1 Regularized Logistic Regression. Su-In Lee, Honglak Lee, Pieter Abbeel and Andrew Y. Ng. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06), 2006.

[2] Y. Nesterov, Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization), 1st ed. Springer Netherlands.

[3] Steve Wright, NIPS Tutorial, 6 December 2010, http://pages.cs.wisc.edu/~swright/nips2010/sjw-nips10.pdf

[4] Nesterov's Optimal Gradient Method.Yao-Liang Yu. University of Alberta. August 6, 2009 http://webdocs.cs.ualberta.ca/~yaoliang/Non-smoothOptimization.pdf

[5] 冨岡亮太（東京大学） 「機械学習における連続最適化の新しいトレンド」 http://www.nec.co.jp/rd/datamining/project/nec_datamining_seminar15.pdf

[6] EE236C - Optimization Methods for Large-Scale Systems (Spring 2011-12) Prof. L. Vandenberghe, UCLA http://www.ee.ucla.edu/~vandenbe/236C/lectures/gradient.pdf

[7] L. Xiao, "Dual averaging method for regularized stochastic learning and online optimization," in Advances in Neural Information Processing Systems 22, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., 2009, pp. 2116-2124.